Analysis and Classification of Human Microbiomes: Detection of Bioindicators and Optimization through Machine Learning

Jonas Coelho Kasmanas[1,2,3], Ulisses Nunes da Rocha[2], Peter F Stadler[3], André Carlos Ponce de Leon Ferreira de Carvalho[1]

[1]Institute of Mathematics and Computer Sciences, University of São Paulo, 13566-590 São Carlos, Brazil

[2]Department of Applied and Environmental Microbiology, Helmholtz Centre for Environmental Research–UFZ, 04318 Leipzig, Germany

[3]Department of Computer Science and Interdisciplinary Centre of Bioinformatics, University of Leipzig, Leipzig, Germany

jonas.kasmanas@usp.br, andre@icmc.usp.br

Abstract. *The human microbiome plays a crucial role in health and disease, yet extracting meaningful insights from it remains challenging. This work presents an integrated framework: (1) the HumanMetagenomeDB, standardizing metadata from 69,000 human metagenomes; (2) MuDoGeR, a multi-domain genome recovery tool; (3) an evaluation of genome recovery biases; (4) gSpreadComp, a workflow for bacterial resistance-virulence; and (5) an AutoML-driven platform built on prokaryotic genomes enabling biologically interpretable feature selection. Together, these contributions enable researchers to move from raw metagenomic data to actionable biological insights at scale.*

1. Introduction and Problem Characterization

Discovering that trillions of symbiotic microorganisms host humans has transformed human health. These microbes and their habitat form the microbiome, crucial for health, disease, and development. The gut microbiome is now essential in personalized medicine, as it drives individual health variations. The human microbiome contains roughly 25 Phyla, 2,000–2,600 Genera, and 5,000–5,500 Species [Thomas and Segata 2019]. Yet, sequencing advances continually reveal new diversity, showing we have not fully characterized the microbiome or its impact on homeostasis [Pasolli et al. 2019].

Concurrently, metagenomics has revolutionized microbial research by resolving the genetic composition without culturing. However, rapid data growth presents challenges. Metagenomic data availability is inconsistent, often lacking metadata [Kasmanas et al. 2021], which hinders comparative studies. Metagenome-Assembled Genomes (MAGs) offer a powerful alternative to reference-based methods by reconstructing genomes of uncultured organisms, revealing their functional potential in natural habitats [Pasolli et al. 2019]. Yet, recovering high-quality MAGs demands complex pipelines, creating barriers for non-experts.

Additionally, annotated metagenomic datasets are complex and high-dimensional, posing challenges for traditional statistics. Machine learning and AutoML can handle this complexity and uncover subtle, biologically significant patterns [Wang et al. 2019].

These challenges highlight a critical need for integrated approaches to translate metagenomic data into insights at scale.

2. Objectives

This work enables large-scale microbiome meta-studies by solving common challenges in genome-centric analysis. Specific objectives include: (i) creating a centralized, searchable database with standardized metadata for public human metagenomic samples; (ii) developing a streamlined multi-domain genome recovery pipeline; (iii) evaluating MAG recovery biases regarding sequencing depth, species abundance, and taxonomic relatedness; (iv) designing a modular comparative genomics workflow targeting antimicrobial resistance genes (ARGs), virulence factors (VFs), and plasmid-mediated transfer; and (v) building an interactive AutoML platform with a large-scale MAG database to identify functional-potential-based bioindicators of environmental contexts).

3. Contributions

3.1. HumanMetagenomeDB

Although public repositories such as the SRA and MG-RAST host vast metagenomic data, misannotated and unstandardized metadata hinder sample mining and reanalysis. HumanMetagenomeDB addresses this by a curated, standardized interface for searching human whole-genome shotgun samples [Kasmanas et al. 2021]. It harmonizes metadata from SRA and MG-RAST, covering over 69,822 samples and 203 attributes. Standardized attributes span host characteristics (age, sex, BMI), 58 diagnoses, 58 countries, 9 body sites, and sequencing parameters. The interface offers Quick search, Advanced search, and Interactive Map for filtering and geographical selection (Figure 1). A download tool retrieves raw data from source repositories.

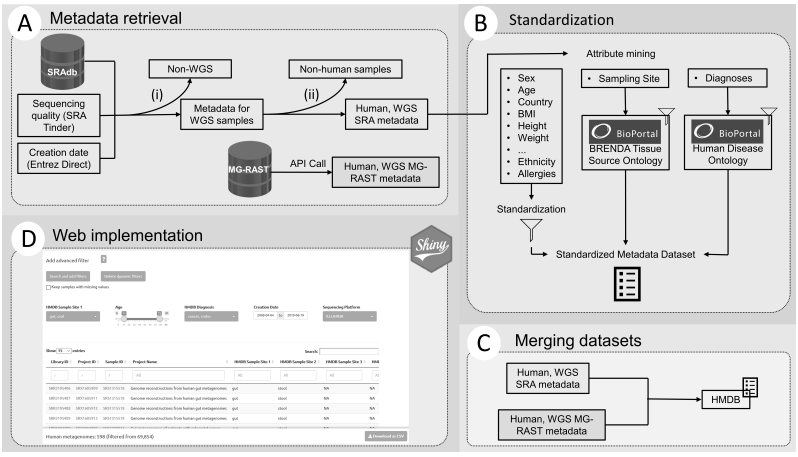

Figure 1. Overview of the HumanMetagenomeDB construction method. (A) metadata retrieval (B) Attribute mining and Standardization. (C) Merging the standardized datasets. (D) The HumanMetagenomeDB web implementation.

3.2. MuDoGeR: Multi-Domain Genome Recovery

After selecting metagenome samples, the next step in a genome-centric study is to recover the MAGs. The MuDoGeR (Multi-Domain Genome Recovery) simplifies this complex

process by integrating established software for assembly, binning, quality assessment, and taxonomic classification into a streamlined workflow [Rocha et al. 2024a]. The tool is divided into five modules (Figure 2): Module 1 preprocesses raw sequences; Module 2 recovers prokaryotic MAGs; Module 3 generates uncultivated virus genomes (UViGs); Module 4 retrieves eukaryotic metagenome-assembled bins (MABs); and Module 5 calculates coverage and relative abundances. Each module generates comprehensive, ready-to-use outputs and can be run independently or in combination. Part of the effort to make MAGs recovery accessible involves automated management of 24 independent software environments and 13 databases, ensuring compatibility and reproducibility. MuDoGeR is distributed as both source code and a Singularity container, reducing installation complexity.

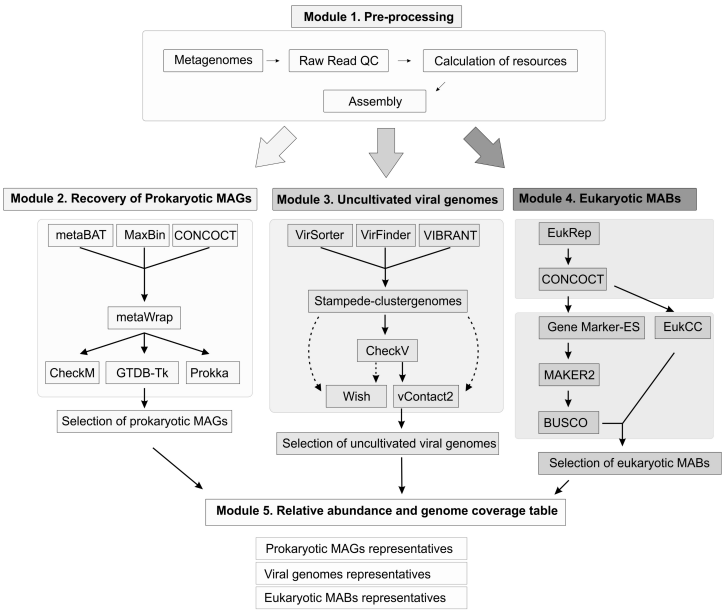

Figure 2. MuDoGeR workflow and outputs.

3.3. Systematic Evaluation of MAG Recovery Biases

Assertions about microbial community composition and functional profiles depend on the reliability of MAG recovery, and systematic errors during recovery can profoundly affect study conclusions [Rocha et al. 2024b]. We conducted a simulation study examining how species abundance, sequencing depth, and taxonomic relatedness influence MAG recovery. Three simulated communities with varying parameters were analyzed using three binning pipelines. Key findings: (1) sequencing depth significantly influences recovery, with an optimal threshold around 60 million reads; (2) high taxonomic relatedness substantially impacts recovery of closely related species; and (3) species abundance influences recovery but is generally less impactful than sequencing depth or taxonomic proximity. These results guide experimental design and interpretation of genome-resolved analyses.

3.4. gSpreadComp

SpreadComp reduces the barrier to comparative genomics by integrating genome annotation, normalization, and sequence comparison into a unified workflow

[Kasmanas et al. 2025]. It integrates six modules: taxonomy assignment, genome quality estimation, ARG annotation, plasmid/chromosome classification, VF annotation, and downstream analysis, which covers gene spread estimation, plasmid-mediated horizontal gene transfer annotation, and resistance-virulence risk ranking. Applied to human gut MAGs across five dietary patterns (Ancient, Ketogenic, Omnivore, Vegan, Vegetarian), it identified differential ARG prevalence and plasmid-mediated horizontal gene transfer potential across conditions. Vegans and Vegetarians showed significantly more ARGs and VFs in plasmid-mediated transfer events than Omnivores and Ketogenic; Bacteroidota from the Ketogenic diet carried statistically more unique VFs than all other diets. These findings illustrate how gSpreadComp accelerates hypothesis generation from in silico analysis to experimental validation.

3.5. Large-Scale MAGs Database and AutoML Platform

The culminating contribution integrates all previous developments into a resource of 514,932 prokaryotic MAGs with rich metadata, including 299,545 newly recovered genomes. Analysis revealed 6,763 bacterial and 31 archaeal species across body sites, with significant body-site specificity. The gut harbored the highest diversity (5,578 species), followed by the oral cavity (1,022 species). 90.4% of gut species were found exclusively at this site. Species accumulation analysis suggested the gut microbiome approaches saturation under current sequencing efforts, while other body sites (skin, airways, and vaginal) remain undersampled.

Finally, we developed an interactive application combining analytical capabilities with visualization. The platform encodes MAG functional potential—from a binary presence/absence matrix of 40,424 non-redundant genes—into a 64-dimensional representation, enabling identification of distinct functional profiles associated with host health, disease, or geographical context. Combined with density-based clustering and AutoML-driven feature selection, it identifies taxonomic groups carrying functionally divergent genomes across conditions—providing biologically interpretable bioindicators. We compared gut MAGs from the USA and China, the two best-represented countries in our dataset. It identified six genera — *Collinsella*, *Acetatifactor*, *Agathobacter*, *Odoribacter*, *Ruminococcus_D*, and *Lachnospira* — as geographical indicators with substantial functional divergence between populations.

4. Results and Discussion

HumanMetagenomeDB has been adopted by the research community and incorporated into NFDI4Microbiota, a German national initiative making microbiological data FAIR (Findable, Accessible, Interoperable, Reusable). It enabled systematic sample selection for all subsequent analyses and promoted metadata standardization in microbiome research.

MuDoGeR scaled to process over 12,000 metagenomic runs, recovering 299,545 new prokaryotic MAGs. It reduced technical barriers, enabling consistent recovery across diverse studies. The simulation study contextualized these results, establishing the 60-million-read optimal threshold and flagging closely related species as a recovery challenge for downstream analysis. Following, applied to dietary groups, gSpreadComp revealed biologically significant patterns. Diets showed distinct ARG and VF profiles, with

diets higher in raw products exhibiting stronger plasmid-mediated horizontal gene transfer potential. Although these findings require experimental validation, they illustrate how streamlined comparative analysis transforms raw genomic data into testable hypotheses.

The 514,932-genome database is one of the most comprehensive collections of human-associated prokaryotic MAGs with standardized metadata. Autoencoder-based functional encoding, combined with AutoML-driven feature selection, provided a novel approach to identifying biologically meaningful patterns. The China-USA use case demonstrated the identification of functionally significant geographical patterns. For instance, *Agathobacter rectalis*, a butyrate-producing bacterium crucial for gut homeostasis, showed distinct functional patterns between Chinese and USA MAGs, potentially reflecting dietary differences, such as the consumption of fermented foods.

Collectively, these contributions form an end-to-end framework advancing the accessibility and interpretability of genome-resolved metagenomic analysis. By reducing technical barriers—from data discovery to biological interpretation—this work enables a broader research community to leverage metagenomic data for understanding the human microbiome.

References

Kasmanas, J. C., Bartholomäus, A., Corrêa, F. B., Tal, T., Jehmlich, N., Herberth, G., von Bergen, M., Stadler, P. F., Carvalho, A. C. P. d. L. F. d., and Nunes da Rocha, U. (2021). HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. *Nucleic Acids Research*, 49(D1):D743–D750.

Kasmanas, J. C., Magnúsdóttir, S., Zhang, J., Smalla, K., Schloter, M., Stadler, P. F., de Leon Ferreira de Carvalho, A. C. P., and Rocha, U. (2025). Integrating comparative genomics and risk classification by assessing virulence, antimicrobial resistance, and plasmid spread in microbial communities with gspreadcomp. *GigaScience*, 14:giaf072.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M. C., Rice, B. L., DuLong, C., Morgan, X. C., Golden, C. D., Quince, C., Huttenhower, C., and Segata, N. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, 176(3):649–662.

Rocha, U., Kasmanas, J. C., Kallies, R., Saraiva, J. P., Toscan, R. B., Štefanič, P., Bicalho, M. F., Borim Correa, F., Baştürk, M. N., Fousekis, E., Viana Barbosa, L. M., Plewka, J., Probst, A. J., Baldrian, P., and Stadler, P. F. (2024a). MuDoGeR: Multi-Domain Genome recovery from metagenomes made easy. *Molecular Ecology Resources*, 24(2):e13904.

Rocha, U., Kasmanas, J. C., Toscan, R., Sanches, D. S., Magnusdottir, S., and Saraiva, J. P. (2024b). Simulation of 69 microbial communities indicates sequencing depth and false positives are major drivers of bias in prokaryotic metagenome-assembled genome recovery. *PLOS Computational Biology*, 20(10):e1012530.

Thomas, A. M. and Segata, N. (2019). Multiple levels of the unknown in microbiome research. *BMC Biology*, 17(1):48.

Wang, C., Wu, Q., Weimer, M., and Zhu, E. (2019). FLAML: A Fast and Lightweight AutoML Library. Version Number: 3.